A quasi-Monte Carlo Metropolis algorithm

Art B. Owen[†] and Seth D. Tribble

Department of Statistics, Stanford University, Stanford, CA 94305

Edited by David O. Siegmund, Stanford University, Stanford, CA, and approved April 11, 2005 (received for review December 21, 2004)

This work presents a version of the Metropolis–Hastings algorithm using quasi-Monte Carlo inputs. We prove that the method yields consistent estimates in some problems with finite state spaces and completely uniformly distributed inputs. In some numerical examples, the proposed method is much more accurate than ordinary Metropolis–Hastings sampling.

completely uniformly distributed | Gibbs sampler | low discrepancy | Markov chain Monte Carlo | randomized quasi-Monte Carlo

Monte Carlo simulation methods are widely used in science, engineering, finance, industry, and statistical inference. Recent decades have seen many improvements in Monte Carlo (MC) methods. Much of the progress has been in quasi-MC (QMC) sampling and in Markov chain MC (MCMC). QMC methods improve the accuracy of MC, from a root mean square error of $O(n^{-1/2})$ using n samples to $O(n^{-1+\varepsilon})$ for any $\varepsilon > 0$, or even $O(n^{-3/2+\varepsilon})$ in some settings, for randomized QMC (RQMC). MCMC greatly extends the range of problems that can be handled by MC. It is thus of interest to combine QMC and MCMC. These subjects both have large bodies of literature, but their published intersection is conspicuously small.

In this work, we prove that some, though not all, QMC methods can yield consistent estimators in Metropolis–Hastings MCMC. The QMC constructions that can be made to work are ones that are "completely uniformly distributed" (CUD) as described below. Using such a QMC construction is similar to using the entire period of a (small) random number generator (RNG). In numerical investigations, QMC can bring a dramatic improvement over MC in some examples and no improvement in others. In the numerical examples we tried, our hybrid of QMC and MCMC always reduced the variance, sometimes by a factor of >200.

This work is organized into the following sections. *Background* gives our notation and some background information on MC, QMC, and MCMC. *A Hybrid of QMC and MCMC* describes CUD sequences and presents our hybrid method, using CUD points for proposals and acceptance in the Metropolis–Hastings algorithm. *Consistency* gives sufficient conditions under which the hybrid yields consistent estimates. *Gibbs Sampler* describes how to fit Gibbs sampling into the framework of this work. *Illustration* has some numerical examples. *Conclusions* states our findings. We finish this section by describing related prior work.

The absence of a QMC approach for the Metropolis algorithm was noted in ref. 1 and again in the recent dissertation by Chaudary (2). Ostland and Yu (3) propose a manually adaptive QMC as an alternative to the Metropolis algorithm. Liao (4) published a proposal for using QMC points in MCMC sampling. He runs a Gibbs sampler using proposals built from a list of n QMC points assembled in a randomized order. He reports an empirical variance reduction but notes that there is no mathematical justification for his procedure. Reordering of quasi-random heat particles (5) between steps has been shown to work for simulation of kinetic equations, but the structure of that problem is different from that of Liao. Particle filters using QMC are discussed in ref. 6. Chaudary (2) uses QMC for the proposal step of a modified Metropolis algorithm that weights rejected proposals. The result was improved accuracy for some numerical examples and essentially unchanged accuracy for others but no mathematical justification.

Our inspiration for looking at these sequences arises from recent work viewing the entire period of a RNG as a QMC rule, a possibility suggested by ref. 7. That technique has been tried on finite dimensional quadrature problems using congruential generators (8) and shift register (Tausworthe) generators (9). MCMC requires simulation of a process that typically uses infinite dimensional inputs. An infinite dimensional ruin process of an insurance company is simulated in ref. 10 using the whole period of a small congruential generator. They report a variance reduction but provide no mathematical justification.

Our proposed hybrid uses QMC within one or more simulated Markov chains. It is also possible to use variance reduction methods, similar to QMC, between two (11) or more (12) chains, where different chains have antithetically coupled movements.

Background

We suppose that the reader is already familiar with simple MC, which we briefly outline here. Then we introduce QMC, RQMC, and MCMC. For a full exposition, see ref. 13 for MC, ref. 14 for QMC, ref. 15 for RQMC, and ref. 16 for MCMC.

MC. In simple MC, a quantity μ of interest is expressed as $\mu = E(f(X))$ for a real valued function f of a random vector X with distribution p. Often p is a probability density on $\mathbb{R}^d$ and then μ is the integral $\int_{\mathbb{R}^d} f(x)p(x)dx$. In other settings p may be a probability mass function. In simple MC, one employs independent random vectors $x_i = (x_{i1}, \ldots, x_{id}) \sim p$ for $i = 1, \ldots, n$ and then estimates μ by $\hat{\mu}_n = (1/n)\Sigma_{i=1}^n f(x_i)$. The justification for simple MC is the law of large numbers. If $E(f(X)^2) < \infty$, then the root mean square error for MC is $O(n^{-1/2})$, and asymptotic confidence intervals are available by the central limit theorem.

The p distributed random vectors x_i are usually computed by transformations of d or more independent uniformly distributed random variables (17). Typically, one uses imperfect but well-tested pseudo-random numbers to simulate the underlying uniform random numbers.

QMC. The focus in QMC sampling is integration over the unit cube. QMC is applicable when one can rearrange the problem so that x_i has the $U[0, 1]^d$ distribution, perhaps changing the value of d in the process. Usually d is finite, though some methods of coping with infinite dimension are given in ref. 18. As with MC, $\hat{\mu}_n$ takes the form $(1/n)\Sigma_{i=1}^n f(x_i)$, but now the x_i values are carefully chosen deterministic points in $[0, 1]^d$.

In QMC, the points x_i are arranged to be more uniformly distributed than random points would be. Their degree of

This paper was submitted directly (Track II) to the PNAS office.

Freely available online through the PNAS open access option.

Abbreviations: CUD, completely uniformly distributed; MC, Monte Carlo; MCMC, Markov chain MC; QMC, quasi-MC; RQMC, randomized QMC; MSE, mean squared error; RNG, random number generator.

[†]To whom correspondence should be addressed at: Department of Statistics, 390 Serra Mall, Stanford, CA 94305. E-mail: owen@stat.stanford.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0409596102

uniformity typically is quantified as a distance between the discrete uniform distribution on x_i and the continuous uniform distribution on $[0, 1]^d$. The most prominent such distance is the star discrepancy, a generalization of the Kolmogorov–Smirnov distance. To define the star discrepancy, first let $\delta(a) = \mathrm{Vol}([0, a]) - (1/n)\Sigma_{i=1}^{n} 1_{x_i \in [0, a]}$ be the local discrepancy function at the point $a \in [0, 1]^d$. Here $\mathrm{Vol}(S)$ is the d-dimensional volume of the (measurable) set S, and $[0, a]$ denotes a d-dimensional box with 0 and a at opposite corners. The star discrepancy is

$$D_n^* = D_n^*(x_1, \ldots, x_n) = \sup_{a \in [0, 1]^d} |\delta(a)|.$$

When $D_n^* \to 0$, then $\hat{\mu}_n \to \mu$ for Riemann integrable f, providing a deterministic version of the law of large numbers for QMC.

The significance of star discrepancy arises from the Koksma–Hlawka inequality

$$|\hat{\mu}_n - \mu| \le D_n^*(x_1, \ldots, x_n) \|f\|_{\mathrm{HK}}, \qquad [1]$$

where $\|f\|_{\mathrm{HK}}$ is the d-dimensional total variation of f in the sense of Hardy and Krause. There are many alternative discrepancies for x_i, and corresponding norms on f, for which a bound like Eq. **1** holds (19).

Widely used QMC points satisfy $D_n^* = O(n^{-1} \log(n)^{d-1})$ as $n \to \infty$. Thus, the error in QMC is $O(n^{-1+\varepsilon})$ for any $\varepsilon > 0$. This rate of convergence is superior to that for MC. The rate is slow to take hold, but empirical comparisons often find that QMC outperforms MC for reasonable n and seldom find QMC to be worse than MC.

To fix ideas, we describe some QMC sequences. Let the integer $n \ge 0$ be written as $n = \Sigma_{k=1}^{\infty} a_{nkb} b^{k-1}$ for an integer base $b \ge 2$ and nonnegative integers $a_{nkb} < b$. The radical inverse function $\phi_b(n) = \Sigma_{k=1}^{\infty} a_{nkb} b^{-k}$ "reflects" the base b expansion of n through the decimal point. The van der Corput sequence has $x_i = \phi_2(i) \in [0, 1]$. Halton's sequence has $x_i = (\phi_{p1}(i), \ldots, \phi_{p_d}(i)) \in [0, 1]^d$, where the p_j are relatively prime. Usually p_j is the jth prime.

Lattice rules (20) are another form of QMC sequence. For a positive integer N and a vector $g = (1, g_1, \ldots, g_{d-1})$ of integers, the lattice rule has $x_i = ig/N - \lfloor ig/N \rfloor$ componentwise for $i = 1, \ldots, N$, where $\lfloor z \rfloor$ is the greatest integer less than or equal to z. A special case are Korobov rules where $g_j = a^{j-1}$ (modulo N) for carefully chosen integers a and N with $1 < a < N$. The Korobov points are related to the points of a multiplicative congruential RNG with $r_i = ar_{i-1}$ mod N. Commonly N is a prime number and a a primitive element modulo N. In this case the RNG has period 1 if started at $r_0 = 0$ and period $N - 1$ otherwise. After a reordering, the nonzero Korobov points are the $N - 1$ points $(r_i/N, \ldots, r_{i+d-1}/N)$ for $i = 1, \ldots, N - 1$ and any $r_0 \ne 0$.

If we run through the RNG once, then we use only $(N - 1)/d$ of the possible d-tuples from the Korobov points. To use all of the d-tuples among the Korobov points requires multiple runs through the RNG, taking care not to repeat Korobov points.

RQMC. The Koksma–Hlawka bound in Eq. **1** is poorly suited to error estimation. It contains the discrepancy D_n^*, which can be hard to compute, and the variation $\|f\|_{\mathrm{HK}}$ that is ordinarily harder to find than μ. Also, although the *Inequality 1* holds as an equality for some worst case f, it can be extremely conservative for integrands arising in applications.

RQMC methods are a hybrid of QMC and MC. RQMC points are usually constructed so that, individually, x_i has the $U[0, 1]^d$ distribution, whereas collectively the x_i have low discrepancy, with probability one. RQMC allows error estimation through confidence intervals for μ based on independent replications of the RQMC estimate. A surprising benefit is that some forms of RQMC reduce the root mean square error to $O(n^{-3/2+\varepsilon})$ on suitably smooth integrands, as shown in ref. 21.

A particularly simple form of randomization is Cranley–Patterson rotation (22). The rotated versions of $a_1, \ldots, a_n \in [0, 1]^d$ are $x_i = a_i + U - \lfloor a_i + U \rfloor$ for a rotation vector $U \sim U[0, 1]^d$ common to all n points.

Standard Construction for Markov Chains. For very simple Markov chains on finite state spaces, one can sample by a standard construction based on inversion of the cumulative distribution function. Let Z be a random variable on values ω_k for $1 \le k \le K < \infty$. To sample Z by inversion, define $P_k = \Sigma_{1 \le l \le k} P(Z = \omega_l)$, draw a sample $u \sim U[0, 1]$, and take $Z = \omega_k$ where k is the smallest index with $u \le P_k$.

The standard construction for sampling a Markov chain is as follows. Begin by sampling x_1 by inversion from the stationary distribution p. Then for $i \ge 1$ sample x_{i+1} by inversion using the conditional distribution of x_{i+1} given x_i. This standard construction is used as a mathematical device in our proofs. We do not assume it can be implemented.

MCMC. MCMC is commonly used in problems where it is difficult or virtually impossible to sample x_i independently from p, by inversion or any other method. Instead, one samples x_i dependently from a Markov chain constructed to have p as a stationary distribution.

Metropolis–Hastings algorithms for MCMC work in two stages: proposal and acceptance. Given x_i, a value y_{i+1} is drawn from a proposal distribution. If that proposal is accepted, then $x_{i+1} = y_{i+1}$, and otherwise $x_{i+1} = x_i$. Let $p_i(x \to y)$ denote the probability, or the probability density, of proposing $y_{i+1} = y$ when $x_i = x$. When $y = x$ it is moot whether y is accepted or rejected. For $y \ne x$ the acceptance probability in Metropolis–Hastings is always

$$A_i(x \to y) = \min\left(1, \frac{p(y)p_i(y \to x)}{p(x)p_i(x \to y)}\right). \qquad [2]$$

The term Metropolis–Hastings is used for the generalization by ref. 23 of the Metropolis algorithm in ref. 24.

Where versions of Metropolis–Hastings differ is in the proposal distribution. In the original Metropolis algorithm, the proposed increments $y_{i+1} - x_i$ are independent and identically distributed. In the independence sampler, the proposals y_{i+1} themselves are independent and identically distributed. Sometimes the standard construction can be viewed as Metropolis–Hastings with acceptance probability one. For example, it suffices to have a reversibility condition wherein $p(y)p_i(y \to x) = p(x)p_i(x \to y) > 0$.

In the Gibbs sampler, the proposal y_{i+1} changes at most one of the components of x_i. In one version the changing component $j(i)$ is chosen randomly and in another $j(i)$ repeatedly cycles through the components of x_i in order. In both cases the changing component is sampled from its conditional stationary distribution given the values of all the nonchanging components.

A Metropolis–Hastings algorithm is "homogenous" if the proposal distribution $p_i(x \to y)$ does not depend on the step i. In that case A_i does not depend on i either. All of the proposals described above are homogenous except for the cyclic Gibbs sampler.

Once again $\mu = \int f(x)p(x)\,dx$ is estimated by a sample mean $\hat{\mu}_n = (1/n)\,\Sigma_{i=1}^{n} f(x_i)$, but now we rely on ergodicity to determine when $\hat{\mu}_n$ tends to μ. Sometimes the first few x_i are skipped. Skipping a finite number of x_i does not affect whether $\hat{\mu}_n \to \mu$, and so we ignore it in this work.

A Hybrid of QMC and MCMC

Our QMC–MCMC hybrid generates the proposals and the acceptances in MCMC using QMC points instead of MC points. There are intuitive arguments for and against this proposal.

First, MCMC sampling has a sequential nature that the usual QMC sampling methods do not respect. For example, with van der Corput points $v_i \in [0, 1]$, it is easy to show that $v_{2k} \in [0, 1/2)$ and $v_{2k+1} \in [1/2, 1)$. Clear and even humorous failures will arise from using van der Corput points in MCMC. Morokoff and Caflisch (25) describe an example where a heat particle supposed to undergo a symmetric random walk will instead move only to the left when sampled by van der Corput points.

The argument in favor of using QMC is that one might expect a good result from MCMC if one ran the chain through one complete period of the underlying RNG. Such a strategy essentially would average together many shorter portions of the generator that might have been presumed to be usable. The entire period of an RNG typically has much lower discrepancy than one would see in an independently and identically distributed sample of the same size. Some, but not all, finite QMC sequences look like RNGs with a small period. Those that do approximate CUD sequences as described below.

Definition 1 (CUD): The sequence $u_1, u_2, \cdots \in [0, 1]$ is CUD, if for every integer $d \geq 1$, the points $z_i = (u_i, \ldots, u_{i+d-1}) \in [0, 1]^d$ satisfy $\lim_{n\to\infty} D_n^*(z_1, \ldots, z_n) = 0$.

The concept of CUD sequences originated with Korobov (26) and is used as definition R1 of randomness by Knuth (27). An up-to-date account of CUD sequences, including some new constructions, is in ref. 28. CUD sequences exist in which $D_n^*(z_1, \ldots, z_n) = O(n^{-1+\varepsilon})$ holds for $z_i = (u_i, \ldots, u_{i+d-1}) \in [0, 1]^d$ and for all integers $d \geq 1$. *Definition 1* applies to an infinite sequence, and RNGs with finite state spaces must have finite length, or at least a finite period. Typically, the CUD property applies to a sequence of RNGs of increasing period. See, for example, theorems 7.3 and 7.4 of ref. 14, which show how certain sequences of linear congruential generators approximate CUD sequences. The role of CUD sequences in simulating processes has been noted previously for stochastic differential equations (29).

Definition 1 groups the u_i into overlapping d-tuples. The hybrid we propose in *Consistency* uses nonoverlapping d-tuples. Chentsov (30) notes the following.

Lemma 1. *If the sequence* $u_1, u_2, \cdots \in [0, 1]$ *is CUD and* $z_i = (u_{di-l+1}, \ldots, u_{di-l+d})$ *for integers* $d \geq l \geq 1$, *then*

$$\lim_{n\to\infty} D_n^*(z_1, \ldots, z_n) = 0.$$

Our grouping uses $l = d$. The more general result in *Lemma 1* allows one to skip the first $d - l$ values u_i.

Independent random points $u_i \sim U[0, 1]$ are CUD in the sense of strong convergence: $\Pr(\lim_{n\to\infty} D_n^* = 0) = 1$ for any d, using either blocked or overlapping vectors (see ref. 28). A definition in the sense of weak convergence suffices for our purposes.

Definition 2 (Weakly CUD): The random sequence $u_1, u_2, \cdots \in [0, 1]$ is weakly CUD if

$$\lim_{n\to\infty} \Pr(D_n^*(z_1, \ldots, z_n) > \varepsilon) = 0$$

holds for every integer $d \geq 1$ and every $\varepsilon > 0$, when $z_i = (u_{di+1}, \ldots, u_{di+d})$.

Consistency

Here we show that one can employ CUD sequences in some Metropolis–Hastings samplers and obtain consistency. We consider chains with finite state spaces $\Omega = \{\omega_1, \ldots, \omega_K\}$ and give conditions under which

$$\hat{p}_n(\omega) \equiv \frac{1}{n}\sum_{i=1}^{n} 1_{x_i=\omega} \to p(\omega), \quad \text{as } n \to \infty, \qquad [3]$$

for all states $\omega \in \Omega$. In the finite state space setting $\hat{\mu}_n \to \mu$ follows from Eq. **3** for all bounded f. Consistency for CUD sampling of Markov chains was proved by Chentsov (30), assuming the standard construction.

Theorem 1 (30). *Let* $x_i \in \{\omega_1, \ldots, \omega_K\}$ *for* $i \geq 1$ *be sampled from the standard construction for Markov chains, using a CUD sequence* u_i. *Assume that all* K^2 *transition probabilities are positive. Then the limit (Eq.* **3***) holds.*

Chentsov's proof uses a coupling idea that we will extend to some Metropolis–Hastings samplers. Where the standard construction uses one number to generate the transition, we suppose that Metropolis–Hastings uses d numbers to generate each transition where $1 \leq d < \infty$. We suppose that the proposal y_{i+1} can be written as a function Ψ_i of x_i and $d - 1$ uniformly distributed random variables. The proposal functions we have in mind are from inversion or other transformations, many of which are described in ref. 17. Then one random variable is used to make the acceptance rejection decision. Specifically, for $i = 0, \ldots, n - 1$,

$$y_{i+1} = \Psi_i(x_i, u_{di+1}, \ldots, u_{di+d-1}), \quad \text{and,} \qquad [4]$$

$$x_{i+1} = \begin{cases} y_{i+1}, & u_{di+d} \leq A_i(x_i \to y_{i+1}) \\ x_i, & \text{else,} \end{cases} \qquad [5]$$

for a state $x_i \in \Omega$ and points $u_j \in [0, 1]$. For a homogenous sampler Ψ_i and A_i do not depend on i.

The law of large numbers for MC sampling applies to Lebesgue integrable functions, whereas that for QMC requires Riemann integrable functions. In typical applications, the distinction need not be drawn. One can, however, make mischief by using a well-behaved transformation Ψ_i when all of $(u_{di+1}, \ldots, u_{di-d-1})$ are irrational numbers and setting y_i to some arbitrary value otherwise. To rule out such pathologies, we suppose that the transitions are regular as described below. Recall that a Jordan measurable set is one whose indicator function is Riemann integrable.

Definition 3 (Regular proposals): The proposals are regular if for all $i \geq 0$, $k \in \{1, \ldots, K\}$, and $l \in \{1, \ldots, K\}$, the set $S_{i,k\to l} = \{(u_{di+1}, \ldots, u_{di+d-1}) \,|\, y_{i+1} = \omega_l \text{ if } x_i = \omega_k\} \subseteq [0, 1]^{d-1}$ is Jordan measurable.

By Lebesgue's theorem (ref. 31, Chapter 8.4) a bounded function on a bounded set $A \subset \mathbb{R}^k$ is Riemann integrable if and only if that function (when extended to 0 on $\mathbb{R}^k - A$) is continuous except on a set of measure zero. Indicator functions are of course bounded, as is the domain $[0, 1]^{d-1}$ in *Definition 3*. Accordingly, the proposals are regular if and only if the sets $S_{i,k\to l}$ have a boundary with $d - 1$ dimensional volume zero. We know of no commonly used proposal functions Ψ_i for which $S_{i,k\to l} \subseteq [0, 1]^{d-1}$ has a boundary of positive $d - 1$-dimensional volume.

Regularity extends easily from proposal sets to transition sets, because unions, complements, and tensor products of Jordan measurable sets are again Jordan measurable. For example, the set of $(u_{di+1}, \ldots, u_{di+d})$ such that $x_i = \omega_k$ transitions to $x_{i+1} = \omega_l \neq \omega_k$ is simply $T_{i,k\to l} = S_{i,k\to l} \times [0, A_i(\omega_k \to \omega_l)]$. The set $T_{ik\to k}$ for self-transitions $x_i = x_{i+1} = \omega_k$ is the complement of $\cup_{l\neq k} T_{i,k\to l}$. A multistep transition through r specific states corresponds to a subset of $[0, 1]^{rd}$ equal to the Cartesian product of r transition sets. The set of vectors in $[0, 1]^{rd}$ for which an r-step transition from $x_i = \omega_k$ to $x_{i+r-1} = \omega_l$ takes place is a union of finitely many multistep transitions sets. When we state below

that a set is Jordan measurable, it follows from the reasoning described in this paragraph.

To generalize Chentsov's Theorem, we will need a "home" state that can be visited from any other and a d-dimensional hyperrectangular region that guarantees a return to the home state.

Definition 4 (Home state): A home state $\omega \in \{\omega_1, \ldots, \omega_K\}$ is one for which there is a box $\mathcal{B} = \prod_{j=1}^{d} [a_j, b_j] \subset [0, 1]^d$ of positive volume such that $x_{i+1} = \omega$ whenever $(u_{di+1}, \ldots, u_{d(i+1)}) \in \mathcal{B}$, regardless of x_i.

As an example, consider proposals y_{i+1} taken from a discrete approximation to the d-dimensional Gaussian distribution. Perhaps y_{i+1} is $(\Phi^{-1}(u_{di+1}), \ldots, \Phi^{-1}(u_{di+d-1}))$ rounded to a discrete set, where Φ is the standard Gaussian cumulative distribution function. Then there is a $d-1$ dimensional box of positive volume near $(1, \ldots, 1)$ that always gives the "highest" proposal y_{i+1}. When the acceptance probability has a positive lower bound, then this highest proposal state is a home state.

Theorem 2. *Let $x_0 \in \Omega = \{\omega_1, \ldots, \omega_K\}$. Suppose that for $i \geq 0$ the proposal $y_{i+1} \in \Omega$ is generated from $(u_{di+1}, \ldots, u_{di+d-1})$ by a homogenous regular proposal and that x_{i+1} is given by Eq.* **5**. *If the sequence $(u_i)_{i \geq 1}$ is CUD and Ω contains a home state with box $\mathcal{B}$, then the Limit* **3** *holds.*

Proof: We will couple x_i to some idealized chains of finite length. These chains exist mathematically, but we do not need to be able to sample from them. For integer $m \geq 1$ let $\tilde{x}_{i,m} = (\tilde{x}_{i,m,0}, \ldots, \tilde{x}_{i,m,m}) \in \Omega^{m+1}$ where $\tilde{x}_{i,m,0}$ is sampled from the stationary distribution p by inversion applied to u_{di}, and for $t = 1, \ldots, m$, the transition from $\tilde{x}_{i,m,t-1}$ to $\tilde{x}_{i,m,t}$ is made by the same function of $(u_{di+1}, \ldots, u_{d(i+1)})$ that the Metropolis–Hastings rule uses to find the successor of $\tilde{x}_{i,m,t-1}$.

Fix a state $\omega \in \Omega$ and let $\varepsilon > 0$. Next,

$$\left| p(\omega) - \frac{1}{n}\sum_{i=1}^{n} 1_{x_i=\omega} \right| \leq \left| \frac{1}{n}\sum_{i=1}^{n} (1_{\tilde{x}_{i,m,m}=\omega} - 1_{x_{i+m}=\omega}) \right| + \left| p(\omega) - \frac{1}{n}\sum_{i=1}^{n} 1_{\tilde{x}_{i,m,m}=\omega} \right| + \left| \frac{1}{n}\sum_{i=1}^{n} (1_{x_{i+m}=\omega} - 1_{x_i=\omega}) \right|. \quad [6]$$

The point x_{i+m} can only differ from the coupled point $\tilde{x}_{i,m,m}$ if the box $\mathcal{B}$ leading to the home state is avoided by m consecutive transitions. There is thus a dm-dimensional hyperrectangle $\mathcal{T}_1$ of volume $(1 - \mathrm{Vol}(\mathcal{B}))^m$ such that $\tilde{x}_{i,m,m} = x_{i+m}$ when $(u_{di+1}, \ldots, u_{di+dm}) \notin \mathcal{T}_1$. We choose m large enough that $(1 - \mathrm{Vol}(\mathcal{B}))^m < \varepsilon$. Because u_i is CUD, the limit as $n \to \infty$ of the first term in the bound Eq. **6** is no larger than ε.

For the second term in Eq. **6**, notice that $\tilde{x}_{i,m,m} = \omega$ if and only if $(u_{di+1}, \ldots, u_{di+dm})$ lies in a dm-dimensional region $\mathcal{T}_2$. The region $\mathcal{T}_2$ has volume $p(\omega)$ because the construction of $\tilde{x}_{i,m,m}$ would generate $\tilde{x}_{i,m,m} \sim p$ had $(u_{di+1}, \ldots, u_{di+dm})$ been sampled $U[0, 1]^{dm}$. For regular proposals, the indicator function of $\mathcal{T}_2$ is Riemann integrable. Because u_i are CUD, it follows that $|p(\omega) - 1/n \sum_{i=1}^{n} 1_{\tilde{x}_{i,m,m}=\omega}| \to 0$ as $n \to \infty$.

The third term in Eq. **6** is no larger than $m/n \to 0$. Therefore the limit, as $n \to \infty$, of the bound in Eq. **6** is no larger than ε. □

Chentsov's *Theorem 1* requires that every transition probability is positive, whereas our *Theorem 2* requires a home state. Chentsov's *Theorem 2* requires neither assumption but requires sampling by the standard construction. His *Theorem 2* contains his *Theorem 1*, but they use different techniques, and he remarks that the first one might extend more easily to continuous state spaces. Our *Theorem 3* below extends his *Theorem 2* from chains sampled by the standard construction to chains sampled by Metropolis–Hastings. We also cover weakly CUD points. If a regular and homogenous Metropolis–Hastings sampler is consistent for independently and identically distributed u_i, then it is also consistent for u_i that are CUD or weakly CUD:

Theorem 3. *Let $x_0 \in \Omega = \{\omega_1, \ldots, \omega_K\}$. Suppose that for $i \geq 0$ the proposal $y_{i+1} \in \Omega$ is generated $(u_{di+1}, \ldots, u_{di+d-1})$ by a homogenous regular proposal and that x_{i+1} is given by Eq.* **5**. *Assume that*

$$\lim_{n \to \infty} \Pr(|\hat{p}_n(\omega_k) - p(\omega_k)| > \varepsilon | x_0 = \omega_l) = 0, \quad [7]$$

holds for all ω_k, $\omega_l \in \Omega$ and all $\varepsilon > 0$ when u_i are independent $U[0, 1]$ random variables. If u_i are replaced by CUD u_i, then the consistency result **3** *holds. If u_i are replaced by weakly CUD u_i, then Eq.* **7** *still holds.*

Proof: First, we assume that u_i are CUD and then define some sets of consecutive u_i for which poor convergence is seen. Given x_0 the value $\hat{p}_n(\omega_k)$ is a function of $u_1, \ldots, u_{nd}$, though we suppress this dependence to avoid unwieldy notation. With this understanding, pick $\varepsilon > 0$ and let

$$\mathcal{T}_{lkn}(\varepsilon) = \{(u_1, \ldots, u_{nd}) \in [0, 1]^{nd} \mid |\hat{p}_n(\omega_k) - p(\omega_k)| > \varepsilon, x_0 = \omega_l,\}.$$

From Eq. **7** there is a value m such that $\mathrm{Vol}(\mathcal{T}_{lkm}(\varepsilon)) < \varepsilon/K$ for all $l, k \in \{1, \ldots, K\}$. Let $\mathcal{T}_{km}(\varepsilon) = \cup_{l=1}^{K} \mathcal{T}_{lkm}(\varepsilon)$. Because the proposals are regular, the set $\mathcal{T}_{km}(\varepsilon)$ is Jordan measurable. Also $\mathrm{Vol}(\mathcal{T}_{km}(\varepsilon)) < \varepsilon$.

For $i = 1, \ldots, n$ let $Z_i \in \{0, 1\}$ with $Z_i = 1$ if and only if $(u_{d(i-1)+1}, \ldots, u_{d(i+m-1)}) \in \mathcal{T}_{km}(\varepsilon)$. Define $\hat{p}_{i,m}(\omega_k) = (1/m) \sum_{t=0}^{m-1} 1_{x_{i+t}=\omega_k}$. Notice that $|\hat{p}_{i,m}(\omega_k) - p(\omega_k)| \leq \varepsilon$ holds whenever $Z_i = 0$, and regardless of x_{i-1}.

Next we write

$$\hat{p}_n(\omega_k) = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{i,m}(\omega_k) + \frac{1}{n}\sum_{t=1}^{m-1} \frac{t}{m}[1_{x_{m-t}=\omega_k} - 1_{x_{n+m-t}=\omega_k}]. \quad [8]$$

We note that there is a minor error in Chentsov's version of this identity, his equation **19** (30). The second term in Eq. **8** is smaller than m/n. Then,

$$\begin{aligned} |\hat{p}_n(\omega_k) - p(\omega_k)| &< \frac{1}{n}\sum_{i=1}^{n} Z_i |\hat{p}_{i,m}(\omega_k) - p(\omega_k)| \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} (1 - Z_i)|\hat{p}_{i,m}(\omega_k) - p(\omega_k)| + \frac{m}{n} \\ &\leq \frac{1}{n}\sum_{i=1}^{n} Z_i + \varepsilon + \frac{m}{n} \\ &\to \mathrm{Vol}(\mathcal{T}_{km}(\varepsilon)) + \varepsilon \leq 2\varepsilon, \end{aligned}$$

as $n \to \infty$, establishing our result for CUD u_i.

If u_i are weakly CUD then as before $|\hat{p}_n(\omega_k) - p(\omega_k)| < (1/n)\Sigma_{i=1}^n Z_i + \varepsilon + m/n$. Now taking $d = m$ in the definition of weakly CUD sequences yields for $n > m/\varepsilon$ that

$$\Pr(|\hat{p}_n(\omega_k) - p(\omega_k)| > 4\varepsilon) < \Pr\left(\frac{1}{n}\sum_{i=1}^{n} Z_i > 2\varepsilon\right) \to 0. \quad \square$$

Chentsov (30) proves a converse for the standard construction. For the sequence u_i to be suitable for every Markov chain under the standard construction, it must be CUD. For each non-CUD sequence (30) constructs a chain for which that sequence applied to the standard construction fails to be consistent. A converse holds for *Theorem 3*, too. A sequence u_i that is not CUD must fail to properly cover some rectangle R in some dimension d. We can then construct a chain on $\{\omega_1, \omega_2\}$ that samples independently visiting state ω_2 at step i if and only if $(u_{(i-1)d+1}, \ldots, u_{di}) \in R$. So the sequence fails to provide consistent estimates for this constructed chain.

Gibbs Sampler

The Gibbs sampler is slightly different from the other samplers. Minor changes are required to handle it. We outline the details in this section.

In MCMC, we use d for the number of variables u_j needed to generate a transition. Let D be the (finite) number of components of x_i. This D may differ from d. The random scan version of the Gibbs sampler takes the changing components $j(i)$ independent and uniformly distributed on $\{1, \ldots, D\}$. Accordingly, only one random variable is needed to choose $j(i)$. To fit our framework the same number m of random numbers must be required to make the proposal y_{i+1} regardless of $j(i)$. When inversion is used, then $m = 1$. Counting the acceptance variable, $d = m + 2$ for random scan Gibbs. If updating the jth component takes $m_j < \infty$ random variables, then one can take $m = \max_{1\le j\le D} m_j$ and simply ignore $m - m_{j(i)}$ of the u_l values at step i.

Because all proposals are accepted in Gibbs sampling, the values u_{di} for $i \ge 1$ are not even used by the method. Instead of ignoring every dth variable, it is more natural to use the whole sequence in blocks of $d - 1$ values, with $u_{(d-1)i+1}, \ldots, u_{(d-1)(i+1)}$ generating $y_{i+1} = x_{i+1}$. Let $\tilde{u}_i$ be a sequence made by taking consecutive blocks of $d - 1$ u_i values and inserting some value $v_k \in [0, 1]$ between the kth and $k + 1$-st block. The value v_k provides the (ignored) variable that determines acceptance of y_k. If there exists a sequence v_k for which $\tilde{u}_i$ is CUD, then *Theorem 3* applies to the Gibbs sampler. In fact, independent random $v_k \sim U[0, 1]$ yield a weakly CUD sequence $\tilde{u}_i$, and so Eq. **7** holds for Gibbs sampling driven by $\tilde{u}_i$. But $\hat{p}_n$ is not random because it ignores the v_k, and so Eq. **7** implies that the consistency Eq. **3** also holds for random scan Gibbs sampling.

Deterministic scan Gibbs sampling does not have homogenous proposals. There are D different proposal distributions. Commonly the component $j(i)$ with a proposed change in y_{i+1} satisfies $j(i) - 1 = i \bmod D$. Rather than considering a general nonhomogenous sampler, one instead can split the chain x_i into D subchains of the form x_{l+Di} for $i \ge 1$ and $1 \le l \le D$. Each such chain is homogenous, and if each of them is consistent, then so is the original chain.

Illustration

Our consistency results show that as $n \to \infty$, the QMC–MCMC estimate $\hat{\mu}_n$ will converge to μ. They do not indicate whether QMC–MCMC is better than MCMC, either asymptotically as $n \to \infty$ or in finite sample sizes. The asymptotic superiority of QMC over MC is well established for finite dimensional problems with sample size approaching infinity. To study the effect of finite sample sizes, infinite dimensions, and the use of continuous instead of discrete state spaces, we try some small numerical examples.

Table 1. Comparison of QMC and MC for independence sampling and random walk sampling on a small numerical example

	Independence		Random walk	
	Mean	MSE	Mean	MSE
MC	-3.58×10^{-4}	3.44×10^{-5}	-7.90×10^{-4}	6.67×10^{-5}
QMC	-7.50×10^{-6}	3.32×10^{-6}	4.10×10^{-4}	2.52×10^{-5}

QMC reduces the MSE by 10.3 for the independence sampler and by 2.65 for the random walk.

Our first example has for p the $N(0, 1)$ distribution, and we study estimates of $\mu = E(x)$, known to be zero. We consider the independence sampler with proposals $y_i \sim N(0, 2.4^2)$, for which the acceptance rate is ~50%. We also consider a random walk sampler, $y_i \sim N(x_{i-1}, 2.4^2)$. For each proposal type, the MC version used pseudo-random numbers to propose and accept/reject (by means of Eq. **2**) for 65,521 steps. The QMC version used all 65,521 points from the LCG with $N = 65{,}521$ and $a = 17{,}364$ given in ref. 8. They were arranged in order $(0, 0)$, (u_1, u_2), (u_3, u_4), $\ldots$, (u_{65519}, u_{65520}), (u_2, u_3), (u_4, u_5), $\ldots$, (u_{65520}, u_1). We applied Cranley–Patterson rotation to these N pairs. The first element in each pair generates the proposal, and the second generates the accept/reject decision.

Each algorithm was repeated 300 times. The mean and mean squared error taken over the 300 answers are displayed in Table 1. In each case the mean is close to the true answer, zero. The square mean is small compared with the mean squared error (MSE) so that bias is a negligible part of the MSE. QMC achieves a MSE reduction factor of 2.65 for the random walk example and 10.3 for the independence sampler.

Our second example has been used by refs. 4 and 32. It features 10 pumps, of which pump j has failed s_j times in $t_j \times 1{,}000$ h. The statistical model is Poisson with $\Pr(n_j = m) = e^{-\lambda_j t_j}(\lambda_j t_j)^m/m!$. The unknown failure rates $\lambda_j \ge 0$ have a Gamma density proportional to $\lambda_j^\alpha e^{-\beta\lambda_j}$ where $\alpha = 1.802$ is known and $\beta \ge 0$ has prior density proportional to $\beta^{\gamma-1}e^{-\delta\beta}$ where $\gamma = 0.1$ and $\delta = 1$. A table with s_j and t_j appears in ref. 32 along with the formula they used to choose α. The state vector $x = (\beta, \lambda_1, \ldots, \lambda_{10})$ has 11 dimensions.

We used a Gibbs sampler with deterministic cycles. The starting point used the maximum likelihood estimates s_j/t_j for λ_j together with the full conditional mean of β, given the starting λ_j values. The Gibbs sampling was driven by inversion of Gamma CDFs applied to RQMC points, as described in ref. 32. The RQMC points were an 11-dimensional Cranley–Patterson rotation applied to QMC points. The QMC points using $N = 1021$ and $a = 65$ from ref. 8 start as $(0, \ldots, 0)$, $(u_1, \ldots, u_{11})$, $\ldots$, $(u_{1013}, \ldots, u_{1020}, u_1, u_2, u_3)$, the next run through the RNG starts with $(u_4, \ldots, u_{14})$, and so on, until all 1,021

Table 2. Comparison of QMC and MC for pump example

Pump	MC	QMC	Ratio
λ_1	6.71×10^{-7}	3.99×10^{-9}	168.0
λ_2	7.66×10^{-6}	5.61×10^{-8}	136.5
λ_3	1.52×10^{-6}	8.92×10^{-9}	170.1
λ_4	9.79×10^{-7}	4.65×10^{-9}	210.5
λ_5	9.40×10^{-5}	7.25×10^{-7}	129.8
λ_6	1.49×10^{-5}	1.09×10^{-7}	136.1
λ_7	3.31×10^{-4}	8.71×10^{-6}	38.0
λ_8	3.12×10^{-4}	2.25×10^{-5}	13.9
λ_9	3.93×10^{-4}	3.96×10^{-6}	99.3
λ_{10}	1.84×10^{-4}	1.03×10^{-6}	178.9
β	8.68×10^{-4}	1.07×10^{-5}	80.8

Shown are MC and QMC variances and their ratio for the parameters of the pump data model described in the text.

vectors have been used once. Each algorithm was repeated 300 times. Table 2 shows variance reductions between ≈14 and ≈210 for QMC–MCMC.

Conclusions

In this work, we have shown that QMC points can be used in Metropolis–Hastings sampling without inconsistency. The points must be CUD. In our numerical examples, the QMC–MCMC hybrid consistently had smaller variance than MCMC, sometimes by a small amount, sometimes by a factor of hundreds. The largest of these gains are better than those reported in related empirical work (2, 4, 6). A rough assessment of our estimated variance reductions can be obtained from the $F_{300,300}$ distribution. With 300 replicates, estimated variance reduction factors are within a multiplication factor of 1.25 of the true factors ≈95% of the time. In quadrature problems, the largest QMC gains have been found for integrals of lower effective dimensionality (33). It remains to see where MCMC problems might have similar structure. We saw larger gains in the higher-dimensional Gibbs sampling problem than in the low-dimensional problem, possibly because the lower-dimensional problem involved a discontinuity at the acceptance threshold. We conclude by noting that the extra work in implementing MCMC with QMC is very small. One replaces the RNG by another RNG that has a smaller period and then uses the entire period one or more times.

We thank two anonymous reviewers for comments and Pierre L'Ecuyer and Christiane Lemieux for suggesting multiple runs through the RNG. This work was supported by National Science Foundation Grant DMS-0306612.

1. Caflisch, R. E. & Moskowitz, B. (1995) in *Modified Monte Carlo Methods Using Quasi-Random Sequences*, eds. Niederreiter, H. & Shiue, P. J.-S. (Springer, New York), pp. 1–16.
2. Chaudary, S. (2004) Ph.D. thesis (Univ. of California, Los Angeles).
3. Ostland, M. & Yu, B. (1997) *Stat. Comput.* **7,** 217–228.
4. Liao, L. G. (1998) *J. Comput. Graphical Stat.* **7,** 253–266.
5. Lécot, C. (1989) *J. Comput. Appl. Math.* **25,** 237–249.
6. Ormoneit, D., Lemieux, C. & Fleet, D. J. (2001) in *Lattice Particle Filters*, eds. Breese, J. & Koller, D. (Morgan-Kaufman, San Francisco), pp. 395–402.
7. Niederreiter, H. (1986) *Math. Programming Study* **27,** 17–38.
8. Entacher, K., Hellekalek, P. & L'Ecuyer, P. (1999) in *Quasi-Monte Carlo Node Sets from Linear Congruential Generators*, eds. Niederreiter, H. & Spanier, J. (Springer, Berlin), pp. 86–97.
9. L'Ecuyer, P. & Lemieux, C. (1999) in *Proceedings of the 1999 Winter Simulation Contest*, eds. Farrington, P. A., Nembhard, H. B., Sturrock, D. T. & Evans, G. W. (IEEE Press, Piscataway, NJ), pp. 632–639.
10. L'Ecuyer, P. & Lemieux, C. (1999) in *Proceedings of the 1999 European Simulation Multiconference* (Society for Computer Simulation, Warsaw), Vol. 2, pp. 533–537.
11. Frigessi, A., Gäsemyr, J. & Rue, H. H. (2000) *Ann. Stat.* **28,** 1128–1149.
12. Craiu, R. V. & Meng, X.-L. (2005) *Ann. Stat.*, in press.
13. Fishman, G. (1996) *Monte Carlo: Concepts, Algorithms and Applications* (Springer, New York), p. 600.
14. Niederreiter, H. (1992) *Random Number Generation and Quasi-Monte Carlo Methods* (SIAM, Philadelphia).
15. L'Ecuyer, P. & Lemieux, C. (2002) in *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, eds. Dror, M., L'Ecuyer, P. & Szidarovszki, F. (Kluwer Academic, Boston), pp. 419–474.
16. Liu, J. S. (2001) *Monte Carlo Strategies in Scientific Computing* (Springer, New York).
17. Devroye, L. (1986) *Non-Uniform Random Variate Generation* (Springer, New York), p. 843.
18. Owen, A. B. (1998) *ACM Trans. Modeling Comput. Simul.* **8,** 71–102.
19. Hickernell, F. J. (1996) *SIAM J. Numerical Anal.* **33,** 1995–2016, and corrected printing (1997) **34,** 853–866.
20. Sloan, I. H. & Joe, S. (1994) *Lattice Methods for Multiple Integration* (Oxford Science, Oxford).
21. Owen, A. B. (1997) *Ann. Stat.* **25,** 1541–1562.
22. Cranley, R. & Patterson, T. (1976) *SIAM J. Numerical Anal.* **13,** 904–914.
23. Hastings, W. K. (1970) *Biometrika* **57,** 97–109.
24. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953) *J. Chem. Phys.* **21,** 1087–1091.
25. Morokoff, W. & Caflisch, R. E. (1993) *SIAM J. Numerical Anal.* **30,** 1558–1573.
26. Korobov, N. M. (1950) *Izv. Akad. Nauk SSSR Ser. Matematika* **14,** 215–238.
27. Knuth, D. E. (1998) *The Art of Computer Programming* (Addison–Wesley, Reading, MA) 2nd Ed., Vol. 3.
28. Levin, M. (1999) *Int. Math. Res. Not.*, 1231–1251.
29. Hofmann, N. & Mathé, P. (1997) *Math. Comput.* **66,** 573–589.
30. Chentsov, N. (1967) *Comput. Math. Math. Phys.* **7,** 218–2332.
31. Marsden, J. (1974) *Elementary Classical Analysis* (Freeman, San Francisco).
32. Gelfand, A. & Smith, A. (1990) *J. Am. Stat. Assoc.* **85,** 398–409.
33. Caflisch, R. E., Morokoff, W. & Owen, A. B. (1997) *J. Comput. Finance* **1,** 27–46.